Predicting NFL Running Back Yardage Based on Prior Season's Performance

Zachary Hogan | General Assembly | Data Science | Spring 2013

The Problem & Hypothesis

- Predict NFL running back yardage
- Analysis will use past season's data to build a model
- Hypothesis: using running back performance measures as features in a linear model will produce an accurate prediction of future performance

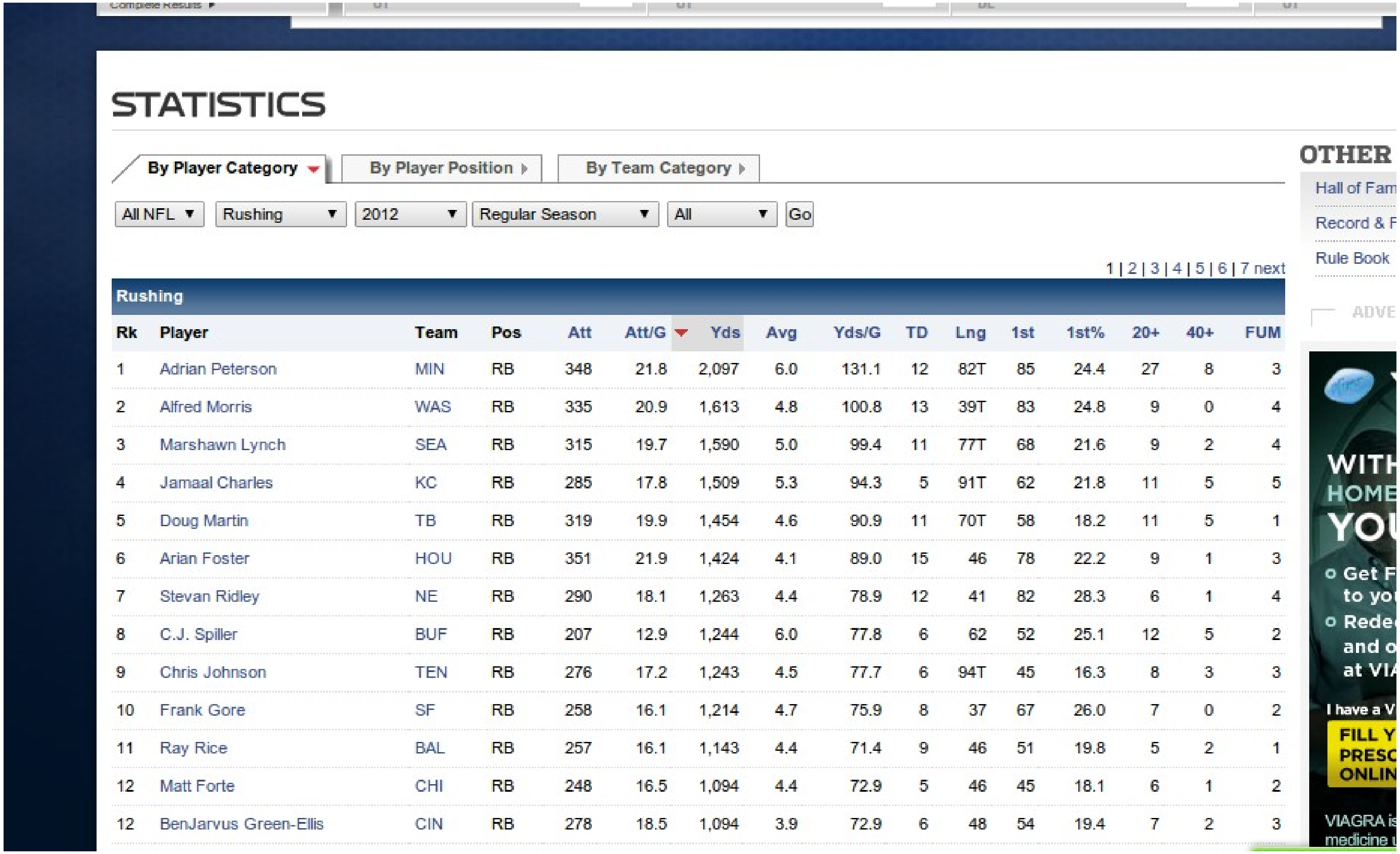
The Data

- Abundance of potential web-sites to collect NFL data
- Not very many unified data sets
- NFL.com most features in one place, trusted source
- 2011 data - training set
- 2010 data - test set
- Will use the model on the 2012 data

The Data – Scraping

- Inspect the HTML
- Data is contained within single table element
- Multiple pages for each year of data
- Store each of these URLs in an array

The Data – Scraping



Complete Results ▸

STATISTICS

Rushing

Rk	Player	Team	Pos	Att	Att/G	Yds	Avg	Yds/G	TD	Lng	1st	1st%	20+	40+	FUM
1	Adrian Peterson	MIN	RB	348	21.8	2,097	6.0	131.1	12	82T	85	24.4	27	8	3
2	Alfred Morris	WAS	RB	335	20.9	1,613	4.8	100.8	13	39T	83	24.8	9	0	4
3	Marshawn Lynch	SEA	RB	315	19.7	1,590	5.0	99.4	11	77T	68	21.6	9	2	4
4	Jamaal Charles	KC	RB	285	17.8	1,509	5.3	94.3	5	91T	62	21.8	11	5	5
5	Doug Martin	TB	RB	319	19.9	1,454	4.6	90.9	11	70T	58	18.2	11	5	1
6	Arian Foster	HOU	RB	351	21.9	1,424	4.1	89.0	15	46	78	22.2	9	1	3
7	Stevan Ridley	NE	RB	290	18.1	1,263	4.4	78.9	12	41	82	28.3	6	1	4
8	C.J. Spiller	BUF	RB	207	12.9	1,244	6.0	77.8	6	62	52	25.1	12	5	2
9	Chris Johnson	TEN	RB	276	17.2	1,243	4.5	77.7	6	94T	45	16.3	8	3	3
10	Frank Gore	SF	RB	258	16.1	1,214	4.7	75.9	8	37	67	26.0	7	0	2
11	Ray Rice	BAL	RB	257	16.1	1,143	4.4	71.4	9	46	51	19.8	5	2	1
12	Matt Forte	CHI	RB	248	16.5	1,094	4.4	72.9	5	46	45	18.1	6	1	2
12	BenJarvus Green-Ellis	CIN	RB	278	18.5	1,094	3.9	72.9	6	48	54	19.4	7	2	3

OTHER

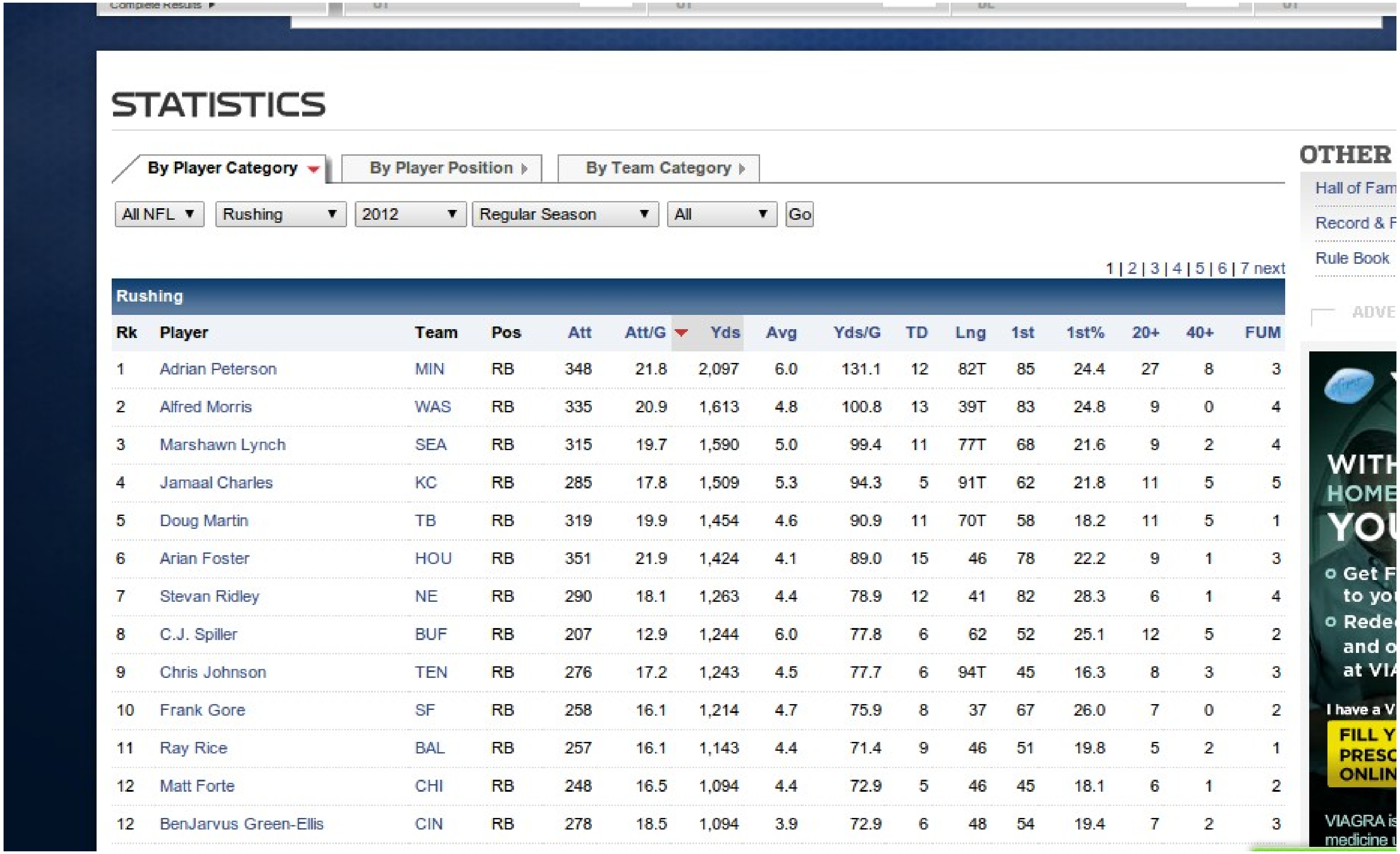
Hall of Fam
Record & F
Rule Book

ADVE
WITH
HOME
YOU
Get F to yo
Rede and o at VIA
I have a V
FILL Y PRESC ONLIN
VIAGRA is medicine

The Data – Scraping

- Inspect the HTML
- Data is contained within single table element
- Multiple pages for each year of data
- Store each of these URLs in an array

The Data – Scraping



Complete Results ▸

STATISTICS

OTHER

- Hall of Fam
- Record & F
- Rule Book

ADVE

WITH HOME YOU
Get F to yo
Rede and o at VIA
I have a V
FILL Y PRESC ONLIN
VIAGRA is medicine

Rushing

Rk	Player	Team	Pos	Att	Att/G	Yds	Avg	Yds/G	TD	Lng	1st	1st%	20+	40+	FUM
1	Adrian Peterson	MIN	RB	348	21.8	2,097	6.0	131.1	12	82T	85	24.4	27	8	3
2	Alfred Morris	WAS	RB	335	20.9	1,613	4.8	100.8	13	39T	83	24.8	9	0	4
3	Marshawn Lynch	SEA	RB	315	19.7	1,590	5.0	99.4	11	77T	68	21.6	9	2	4
4	Jamaal Charles	KC	RB	285	17.8	1,509	5.3	94.3	5	91T	62	21.8	11	5	5
5	Doug Martin	TB	RB	319	19.9	1,454	4.6	90.9	11	70T	58	18.2	11	5	1
6	Arian Foster	HOU	RB	351	21.9	1,424	4.1	89.0	15	46	78	22.2	9	1	3
7	Stevan Ridley	NE	RB	290	18.1	1,263	4.4	78.9	12	41	82	28.3	6	1	4
8	C.J. Spiller	BUF	RB	207	12.9	1,244	6.0	77.8	6	62	52	25.1	12	5	2
9	Chris Johnson	TEN	RB	276	17.2	1,243	4.5	77.7	6	94T	45	16.3	8	3	3
10	Frank Gore	SF	RB	258	16.1	1,214	4.7	75.9	8	37	67	26.0	7	0	2
11	Ray Rice	BAL	RB	257	16.1	1,143	4.4	71.4	9	46	51	19.8	5	2	1
12	Matt Forte	CHI	RB	248	16.5	1,094	4.4	72.9	5	46	45	18.1	6	1	2
12	BenJarvus Green-Ellis	CIN	RB	278	18.5	1,094	3.9	72.9	6	48	54	19.4	7	2	3

The Data – Scraping

- Inspect the HTML
- Data is contained within single table element
- Multiple pages for each year of data
- Store each of these URLs in an array

The Data – Cleaning

- Multiple columns contain numerous escape characters (\n, \t)
- Longest rush column contains 'T' characters
- Commas from the total yardage column
- 2010 test data contains escape characters in additional columns

The Data – Storage

- Merge future season yardage to the test and training sets
- In Python, store the data in Pandas data frame
- Export as csv, to transfer to R or anywhere else we might need it
- Left with a Test set, a Training set, and the data we will predict on

Statistical Method

	Continuous	Categorical
Supervised	regression	classification
Unsupervised	dimension reduction	clustering

Statistical Method

	Continuous	Categorical
Supervised	**REGRESSION**	classification
Unsupervised	dimension reduction	clustering

Statistical Method

- Multivariate Linear Regression
- 2011 data is the training set
- 2010 will be our test set
- If the model performs well on the 2010 data, we will use it to predict the 2013 results based on the 2012 data

Statistical Method

```
train_fit <- lm(yards_2012 ~ att + att.game +
yards + avg + yards.game + TD + long + X1st
+ X20. + X40. + fumbles, data=train_data11)
```

Statistical Method

- This fit produces an R^2 of 0.6006
- Attempt backwards elimination

Statistical Method

```
train_data10["predict_yds"] <-
predict(train_fit, train_data10)
```

Statistical Method
2010 Test Set

Running Back	2011 Predicted Yards	2011 Actual Yards	Difference
Arian Foster	1173	1224	51
Jamaal Charles	564	83	-481
Michael Turner	1131	1340	209
Chris Johnson	1303	1047	-256
Maurice Jones-Drew	793	1606	813
Adrian Peterson	988	970	-18
Rashard Mendenhall	1367	928	-441
Steven Jackson	890	1145	255
Ahmad Bradshaw	988	659	-329
Ray Rice	777	1364	587

Statistical Method
2010 Test Set

- Root Mean Squared Error – 231.8 yards
- Use this to measure to compare the accuracy of future modifications to the model

Statistical Method
2013 Predictions

Running Back	2012 Yards	2013 Predicted Yardage
Adrian Peterson	2097	1700
Alfred Morris	1613	1152
Marshawn Lynch	1590	1022
Jamaal Charles	1509	619
Doug Martin	1454	1161
Arian Foster	1424	1218
Stevan Ridley	1263	670
C.J. Spiller	1244	732
Chris Johnson	1243	746
Frank Gore	1214	743

Business Applications

- NFL team management
- NFL television analysis
- Fantasy Football

Conclusion

- Reasonably accurate prediction on the 2010 test data
- Lots of room for improvement
  - Use only data from top running backs, combine it with multiple years of data
  - Additional features
  - Predict rank rather than yardage

Conclusion

Thank you!